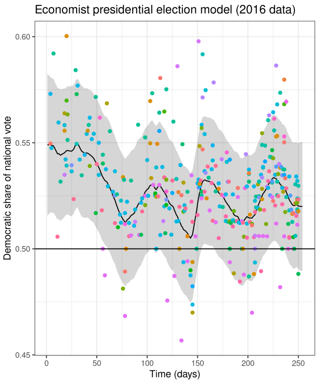
Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

April 2024

Theory and Foundations of Statistics in the Era of Big Data

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

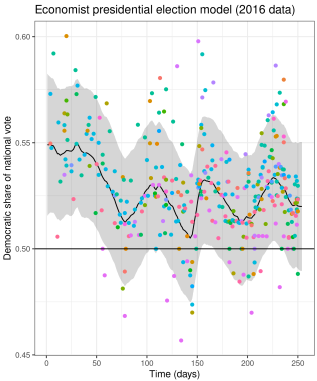
Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

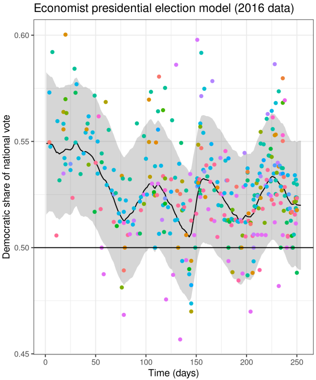
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

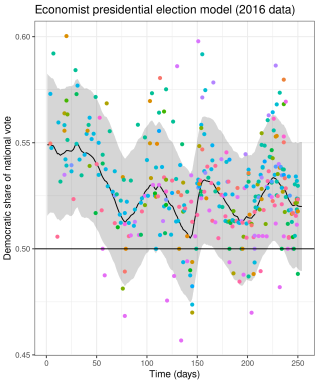
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time
- Is there high variability under re-sampling? [Huggins and Miller, 2023]
  Re-fit with bootstrap samples of data

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

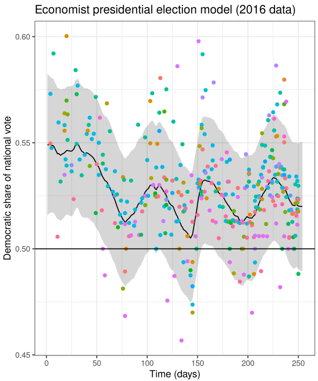
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time
- Is there high variability under re-sampling? [Huggins and Miller, 2023]
  Re-fit with bootstrap samples of data
- Are a small proportion (1%) of polls highly influential? [Broderick et al., 2020]
  Re-fit with sets of all 1% of datapoints removed

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  - Re-fit with data points removed one at a time
- Is there high variability under re-sampling? [Huggins and Miller, 2023]
  - Re-fit with bootstrap samples of data
- Are a small proportion (1%) of polls highly influential? [Broderick et al., 2020]
  - Re-fit with sets of all 1% of datapoints removed

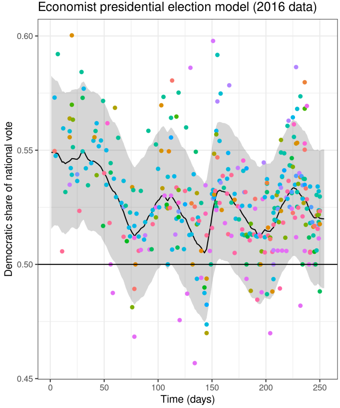
Problem: Each MCMC run takes about 10 hours (Stan, six cores).

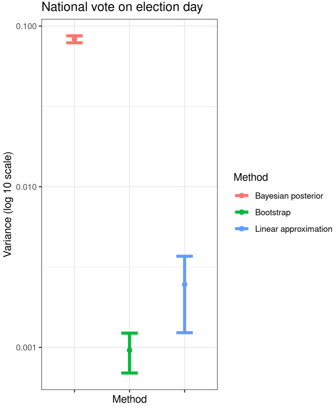
Results

We propose: Use posterior draws based on the full data, to form a linear approximation to *data reweightings*.

Results

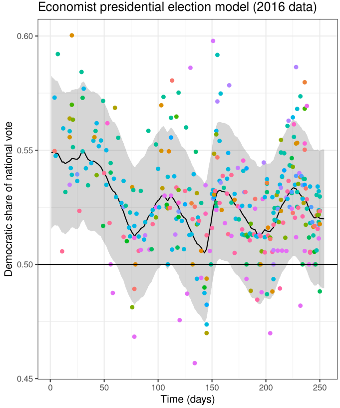
We propose: Use posterior draws based on the full data, to form a linear approximation to *data reweightings*.

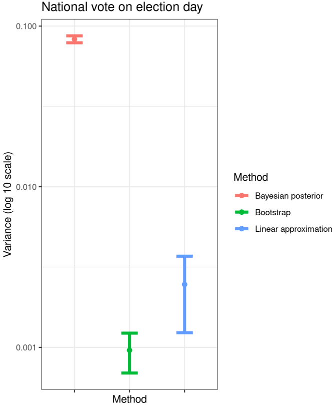




Results

We propose: Use posterior draws based on the full data, to form a linear approximation to *data reweightings*.





Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate, even as $N \to \infty$

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate, even as $N \to \infty$
- A trick question, and some implications of different weightings.

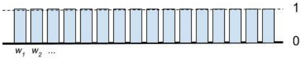
Data re-weighting.

Data re-weighting.

Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:

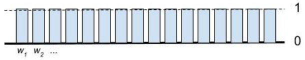


Data re-weighting.

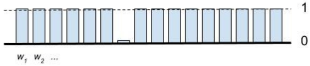
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:



Leave-one-out weights:

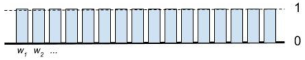


Data re-weighting.

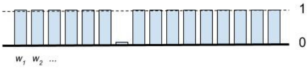
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

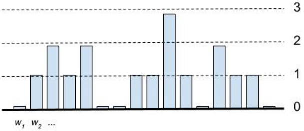
Original weights:



Leave-one-out weights:



Bootstrap weights:

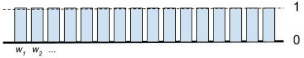


Data re-weighting.

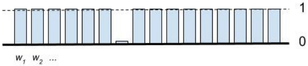
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

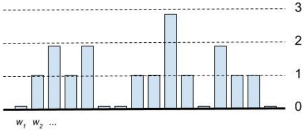
Original weights:

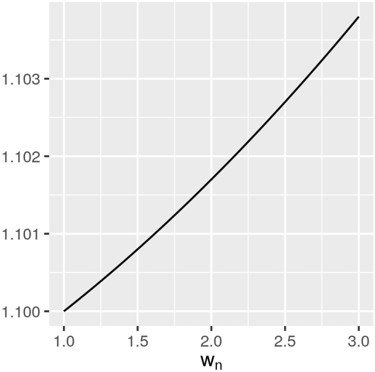


Leave-one-out weights:



Bootstrap weights:



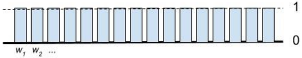


Data re-weighting.

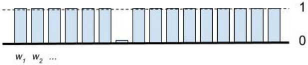
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

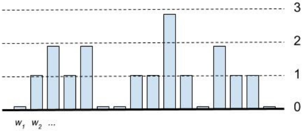
Original weights:

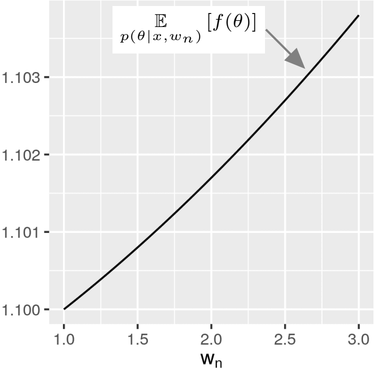


Leave-one-out weights:



Bootstrap weights:



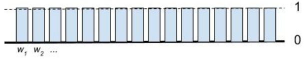


Data re-weighting.

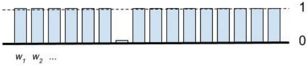
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

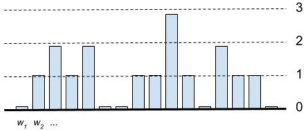
Original weights:

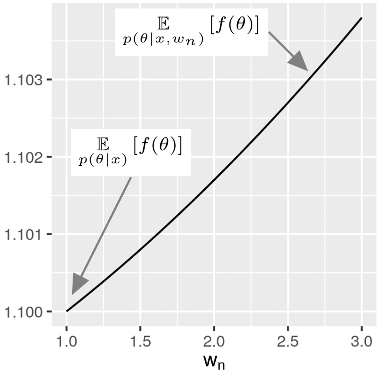


Leave-one-out weights:



Bootstrap weights:



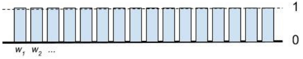


Data re-weighting.

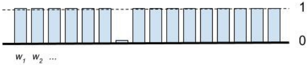
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

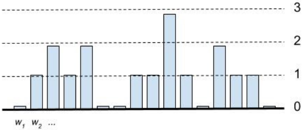
Original weights:

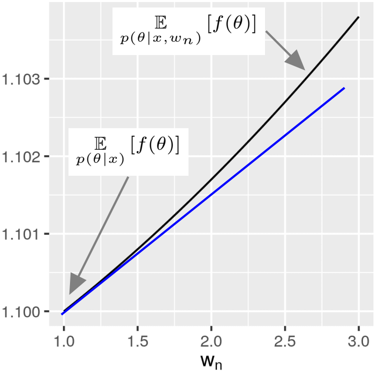


Leave-one-out weights:



Bootstrap weights:



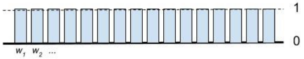


Data re-weighting.

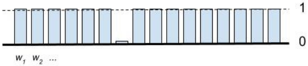
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

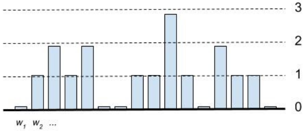
Original weights:

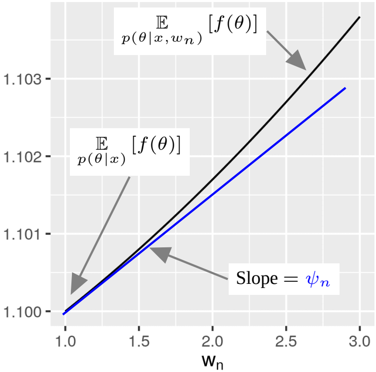


Leave-one-out weights:



Bootstrap weights:



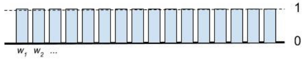


Data re-weighting.

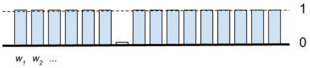
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

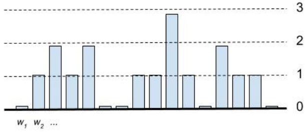
Original weights:

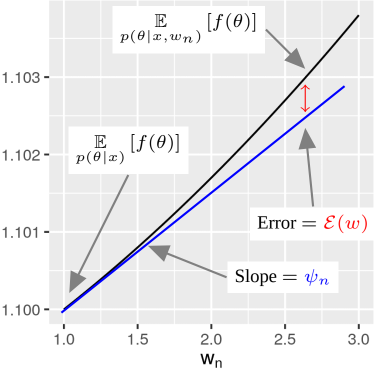


Leave-one-out weights:



Bootstrap weights:



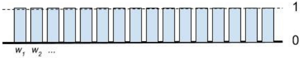


Data re-weighting.

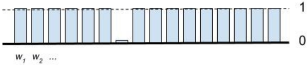
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

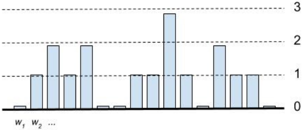
Original weights:

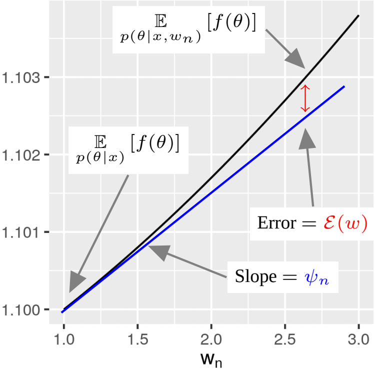


Leave-one-out weights:



Bootstrap weights:





The re-scaled slope $N\psi_n$ is known as the "influence function" at data point x_n.

$$\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})}[f(\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] - \psi_n$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \underset{p(\theta|x,w_{(-n)})}{\mathbb{E}}[f(\theta)] \approx \underset{p(\theta|x)}{\mathbb{E}}[f(\theta)] - \psi_n$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\text{Bootstrap variance} = \underset{p(w)}{\text{Var}} \left(\underset{p(\theta|x,w)}{\mathbb{E}}[f(\theta)] \right) \approx \frac{1}{N^2} \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \underset{p(\theta|x,w_{(-n)})}{\mathbb{E}}[f(\theta)] \approx \underset{p(\theta|x)}{\mathbb{E}}[f(\theta)] - \psi_n$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\text{Bootstrap variance} = \underset{p(w)}{\text{Var}}\left(\underset{p(\theta|x,w)}{\mathbb{E}}[f(\theta)]\right) \approx \frac{1}{N^2}\sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2$$

Influential subsets: Approximate maximum influence perturbation (AMIP).

Let $W_{(-K)}$ denote weights leaving out K points.

$$\max_{w \in W_{(-K)}}\left(\underset{p(\theta|x,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|x)}{\mathbb{E}}[f(\theta)]\right) \approx -\sum_{n=1}^{K}\psi_{(n)}.$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar mean posterior–mean zero (e.g., $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$).

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}(w_n) = \frac{1}{2}\underbrace{\underset{p(\theta|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

$= O_p(N^{-1})$ under a BCLT $\qquad\qquad = O_p(N^{-2})$ under a BCLT

Theorem 2 of Giordano and Broderick [2023] (paraphrase):
If the posterior $p(\theta|X)$ satisfies a kind of Bayesian central limit theorem (BCLT),[a] then the map $w_n \mapsto N\left(\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]\right)$ becomes linear as $N \to \infty$.

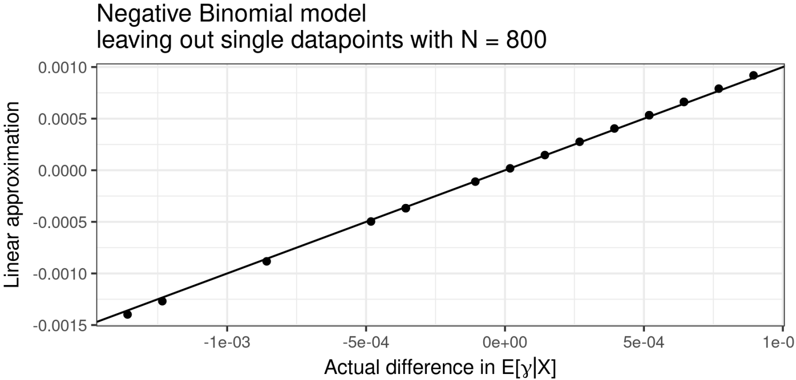
[a] Existing results are sufficient for a *particular weight* [Kass et al., 1990]. Giordano and Broderick [2023] proves a kind of average convergence over all weights.

Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^{N} \mathrm{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

As $N \to \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (a BCLT).

$$\Rightarrow N \left(\underset{p(\gamma|X,w_n)}{\mathbb{E}} [\gamma] - \underset{p(\gamma|X)}{\mathbb{E}} [\gamma] \right) = \psi_n (w_n - 1) + O_p(N^{-1}).$$



High dimensional problems

What about when the posterior doesn't obey a BCLT?

Example: **Poisson model with random effects (REs) λ and fixed effect γ.**

If the observations per random effect remains bounded as $N \to \infty$, then

Parameter λ grows in dimension with N.	Parameter γ is a scalar.
Marginally, $p(\lambda\|X)$ does not concentrate.	Marginally, $p(\gamma\|X)$ obeys a BCLT.

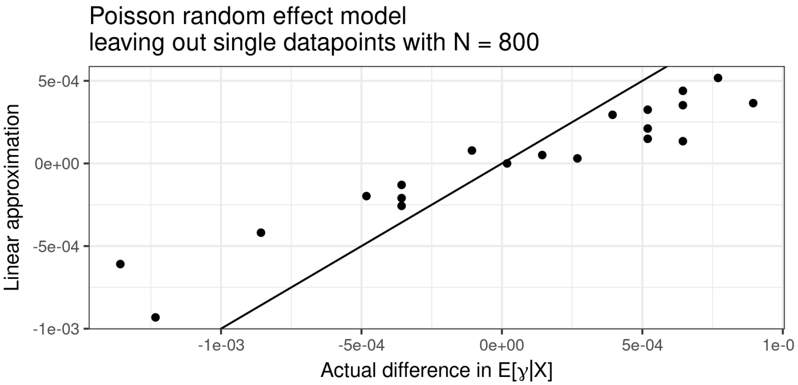
Does $w_n \mapsto \underset{p(\lambda|X,w_n)}{\mathbb{E}}[f(\lambda)]$ become linear as N grows?

Not in general. Since $p(\lambda|X)$ doesn't concentrate, both the slope ψ_n and error $\mathcal{E}(w_n)$ are $O(1)$ in general. $\Rightarrow$ The map $w_n \mapsto \underset{p(\lambda|X,w_n)}{\mathbb{E}}[f(\lambda)]$ is nonlinear in general.

Does $w_n \mapsto \underset{p(\gamma|X,w_n)}{\mathbb{E}}[f(\gamma)]$ become linear as N grows?

Theorem 5 of Giordano and Broderick [2023] (paraphrase):
In the linear approximation to $\underset{p(\gamma|X,w_n)}{\mathbb{E}}[f(\gamma)]$, both the slope ψ_n and the error $\mathcal{E}(w_n)$ are $O_p(N^{-1})$ when $p(\lambda|X,\gamma)$ does not concentrate, even if $p(\gamma|X)$ obeys a BCLT marginally.
In general, **the posterior expectation does not become linear in w_n as N grows.**

Experiments



A contradiction?

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X,w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

$$\log p(X|\gamma, w^m) = \sum_{n=1}^{N} w_n^m \log p(x_n|\gamma)$$

Poisson observations with random effects.

Asymptotically non-linear in w.

$$\log p(X|\gamma, \lambda, w^c) = \sum_{n=1}^{N} w_n^c \log p(x_n|\lambda, \gamma)$$

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X,w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

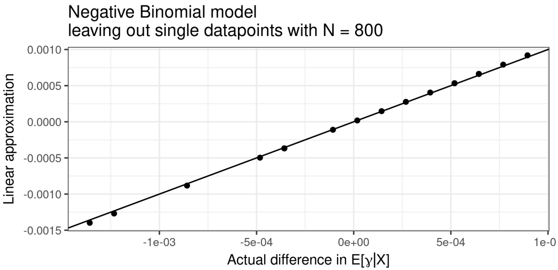
Trick question! We weight a log likelihood contribution, not a datapoint.

The two weightings are not equivalent in general.

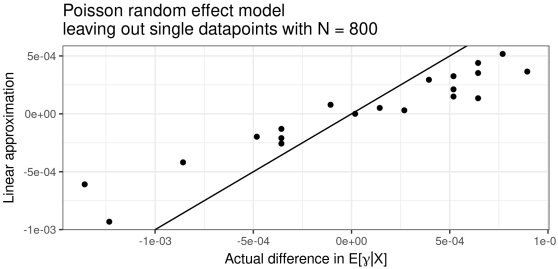
Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Approximation based on $\log p(x_n|\gamma)$.



Approximation based on $\log p(x_n|\gamma, \lambda)$.

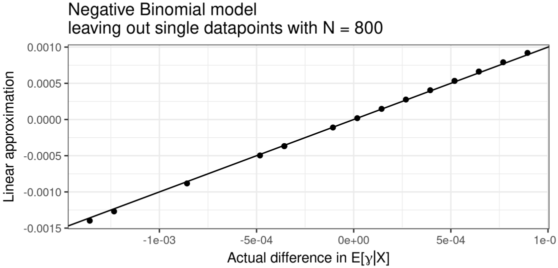


Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

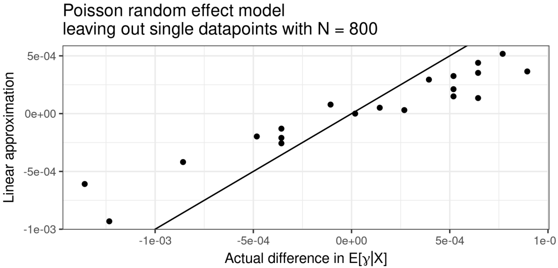
Approximation based on $\log p(x_n|\gamma)$.

Not computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$ in general.



Approximation based on $\log p(x_n|\gamma, \lambda)$.

Computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$.

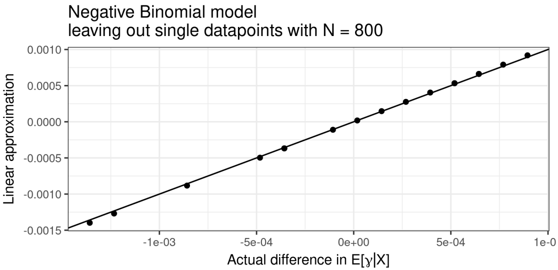


Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Approximation based on $\log p(x_n|\gamma)$.

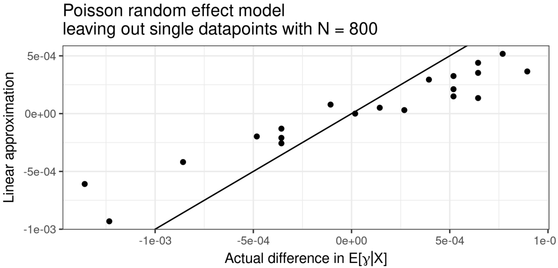
Not computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$ in general.



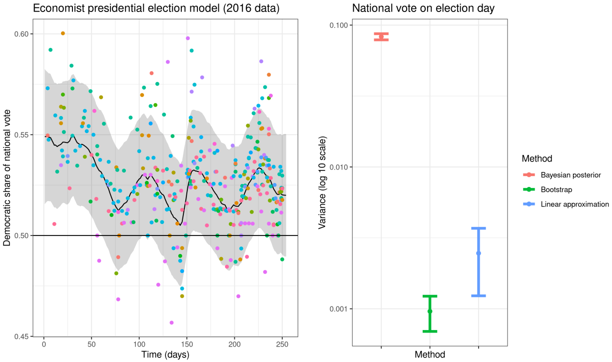
Approximation based on $\log p(x_n|\gamma, \lambda)$.

Computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$.

May still be useful when $p(\lambda|X)$ is *somewhat* concentrated.



Observations and consequences



- We use often use models $p(\gamma, \lambda | X)$, and can't compute $p(\gamma | X)$ analytically.
- There may be multiple ways to define "exchangable unit" in a given problem.

  ... But without nesting, $\log p(x_n | \gamma, \lambda)$ may be the natural model-free exchangeable unit.
- Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.

  ... Especially given the linear approximation's huge computational advantage.

Preprint: Giordano and Broderick [2023] (arXiv:2305.06466)

T. Broderick, R. Giordano, and R. Meager. An automatic finite-sample robustness metric: When can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*, 2020.

A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL `https://projects.economist.com/us-2020-forecast/president`. Data and model accessed Oct., 2020.

R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2023.

J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990.

A. Vehtari and J. Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.